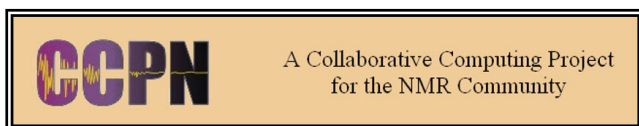


Universal Data Model for Harvesting and Exchange of NMR Results

Software interoperability and data standards are important issues for the efficient use of nuclear magnetic resonance (NMR) data. At present the Protein Data Bank (<http://www.rcsb.org/pdb/>) holds over five thousand structures that were determined by using NMR, and this number is growing steadily every day. In the Protein Data Bank, the relative increase of structures determined by NMR is growing more rapidly than the ones determined by x-Ray methods. NIST is working through the Collaborative Computing Project for the NMR Community (CCPN) to address the critical need to establish data and software standards for NMR data on macromolecules.

T. N. Bhat (Div 831)

The Collaborative Computing Project for the NMR community (CCPN), founded in 1999, is a public non-profit project serving the macromolecular NMR community. CCPN is hosted by the University of Cambridge and The European Bioinformatics Institute.



Through this existing collaboration, NIST held a workshop to discuss a new approach to data and software standards for NMR titled: *A Data Model for Biomolecular NMR Spectroscopy*. Details of that workshop can be found at <http://www.ccpn.ac.uk/index.html>.

The NIST workshop supported the goals of the CCPN, which was established:

- To promote the exchange of ideas through the arrangement of meetings and workshops
- To establish a universal data model for harvesting and exchange of data between different NMR-related software packages, and to develop associated subroutine libraries
- To develop software packages using the common data model to perform standard functions required during the determination of macromolecular structures by NMR
- To promote the use of the common data model and to diffuse NMR-related software made by third parties, especially those that make use of the data model, within the NMR community

NIST has been active in the CCPN and proposed the theory and development of an Application Program Interface (API) for NMR, which was funded by the Biotechnology and Biological Sciences Research Council of the UK. In the last few years, great progress has been made in designing and developing data and software standards for NMR, and can be found on the CCPN website:

http://www.ccpn.ac.uk/project/about_ccpn/about_ccpn.html

The software described here is called *Mempos*. The data model is centered in a UML subset similar to XML. Code is generated automatically for several languages, with Python and Java being supported so far. The product includes an object-oriented data interaction API and its implementation, complete with data validation, and checking and notification facility. Data storage in either XML files or in a relational database is generated in the data access subroutines. XML and database schemas and documentation are also generated from the UML model. Two of the publications that describe this work and the data standards software were published during 2005.

The *Mempos* frame work is a tool for data modeling and for the fully automatic generation of sub-routine libraries for data access in multiple computer languages.

Publications:

Fogh, R. H., Boucher, W., Vranken, W. F., Pajon, A., Stevens, T. J., Bhat, T. N., Westbrook, J., Ionides, J. M. C., Laue, E. "A framework for scientific data modeling and automated software development." *Bioinformatics* (2005) 21,1678-1684.

Fogh, R., Boucher, W., Varken, W., Pajon, A. S, Bhat, T. N., Westbrook, J., Ionides, J., Ernest, L. E. "**Mempos: data modeling and automatic code generation in multiple languages.**" In *Proceeding of the Second European Workshop on Model Driven Architecture with an emphasis on Methodologies and Transformations*, Edited by D. H. Akehurst. Computing Laboratory, University of Kent, Canterbury, Kent, U.K. (2004).